Research Paper | Computer Science | E-ISSN No : 2454-9916 | Volume : 3 | Issue : 10 | Oct 2017

CONCEPTUAL STUDIES OF BIG DATA ANALYSIS

Nagendra Kumar Sahu

Research Scholar, Mats University, Raipur (C.G.)

ABSTRACT

Big Data Analytics may be the key to fighting cyber crime. Using big data to combat cyber crime is becoming a decisive strategy for businesses willing to stay secure. With security risks becoming larger, from structured and unstructured data inside the network servers to smart phones, businesses need to be extremely alert due to tremendous increase in cyber threats. Several organizations are leveraging big data analytics for supporting their business processes. However, there are only few organizations that have realized the potential benefits of analytics towards ensuring information security.

INTRODUCTION:

Big data is a term that is used to describe data that is high volume, high velocity, and/or high variety; requires new technologies and techniques to capture, store, and analyze it; and is used to enhance decision making, provide insight and discovery, and support and optimize processes. Big data analytics is the process of examining large and varied data sets -- i.e., big data -- to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informed business decision Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and information privacy .While big data is the darling of some Science, Technology, Engineering and Mathematics programs, most mainstream high school students have no idea what it is or how it's transforming the business world – and your opportunities for employment. While the industry has only existed for a decade, big data is everywhere

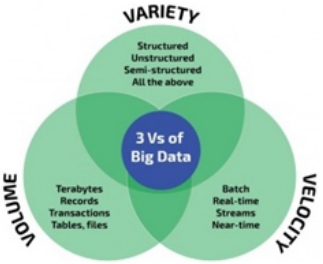

Fig. 1: Big Data Characteristics

TECHNOLOGIES:

2.1 Introduction to MapReduce

The best known example of Big Data execution environment is probably Google MapReduce (Dean, J. and Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Commun. ACM, 51(1):107-113) (the Google's implementation of the MapReduce programming model) and Hadoop, its open source version (Lam, C. (2011). Hadoop in action. Manning, 1st edition). This environment aims at providing elasticity by allowing the adjustment of resources according to the application, handling errors transparently and ensuring the scalability of the system. this programming model is built upon two "simple" abstract functions named Map and Reduce, which are inherited from the classical functional programming paradigms. Users specify the computation in terms of a map (that specify the per-record computation) and a reduce (that specify result aggregation) functions, which meet a few simple requirements. For example, in order to support these, MapReduce requires that the operations performed at the reduce task to be both "associative" and "commutative."

2.2 Hadoop Ecosystem

Apache Hadoop is a Big Data framework that is part of the Apache Software Foundation. Hadoop is an open source software project that is extensively used by some of the biggest organizations in the world for distributed storage and processing of data on a level that is just enormous in terms of volume. That's the reason the Apache Hadoop runs its processing on large computer clusters built on commodity hardware. Some of the features of the Hadoop platform are that it can be efficiently used for data storage, processing, access, analysis, governance, security, operations and deployment. Hadoop is a top level project that is being built and used by a diverse group of developers, users and contributors cutting across nationalities under the auspices of the Apache Foundation. Hadoop is currently governed under the Apache License 2.0. Hadoop operates on thousands of nodes that involve huge amounts of data and hence during such a scenario the failure of a node is a high probability. So the Hadoop platform is resilient in the sense that The Hadoop distributed file systems immediately upon sensing of a node failure divert the data among other nodes thus allowing the whole platform to operate without any interruptions.

The MapReduce framework is based on the fact that most of the information processing tasks consider a similar structure, i.e. the same computation is applied over a large number of records; then, intermediate results are aggregated in some way. As it was previously described, the programmer must specify the Map and Reduce functions within a job. Then, the job usually divides the input dataset into independent subsets that are processed in parallel by the Map tasks. Map Reduce sorts the different outputs of the Map tasks which become the inputs that will be processed by the Reduce task. The main components of this programming model, that were previously illustrated in Fig. 1, are the following ones: Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

METHODOLOGIES:

Confidentiality preserving data mining techniques

Text Data Mining

Textual data represents rich information, but lacks structure and requires specialist techniques to be mined and linked properly as well as to reason with and make useful correlations. A set of techniques will be developed for extracting entities, relations between them, opinions and other elements for use to support semantic indexing and visualization and anonymisation. [Dr Udo Kruschwitz, Professor Massimo Poesio, Professor Maria Fasli, Dr Beatriz de la Iglesia]

1. **Machine learning and transactional data**: Investigate machine learning and other methods for identifying stylised facts, seasonal, spatial or other relations, patterns of behavior at the level of the individual, group, or region from transactional data from business, local government or other organizations. Such methods can provide essential decision support information to organizations in planning services based on predicted trends, spikes or troughs in demand. [Professor Maria Fasli, Dr Beatriz de la Iglesia]

2. **Developing methods to evaluate, target and monitor the provision of care:** Models and statistical methods for the analysis of local government health and social care data will be developed alongside new data mining and machine learning algorithms to identify intervention subgroups, and new joint modelling methods to improve existing predictive models with a view to evaluate, target and monitor the provision of care. [Professor Abdel Salhi, Professor Berthold Lausen, Professor Elena Kulinskaya]

Data quality grading and assurance

This research will develop new and adapt existing methodologies for merging data from multiple sources. It will also develop robust techniques for data quality grading and assurance providing automated data quality and cleaning procedures for use by researchers. [Beatriz de la Iglesia]

3. **Identifying "unusual" data segments:** Methods will be developed to automatically identify "unusual" data segments through an ICMetrics-based technique. Such methods will be able to alert researchers of specific data seg-

ments that require subsequent further analysis and identify potential issues with unsolicited data manipulation and integrity breaches. { Professor Klaus McDonald-Maier}

Some datasets include sensitive information; this research considers how best to aggregate/transform data to allow subsequent analysis to be undertaken with the minimum loss of information. Methods for dimensionality reduction and data perturbation techniques will be investigated alongside privacy preserving data mining methods. [Dr Beatriz de la Iglesia]

Algorithm Used

Clustering algorithms can be used to find relationships within an organization's dataset. These algorithms can be used to find different kinds of groupings within a customer base, or to decide what customers and services can be grouped together. An unsupervised clustering approach can offer some distinct advantages, as compared to the supervised learning approaches. One example is the way novel applications can be discovered by studying how the connections are grouped when a new cluster is formed. Some algorithms are follows:-

- K Means Clustering
- Association Rules
- Linear Regression
- Logistic Regression
- Naïve Bayesian Classifier
- Decision Trees
- Time Series Analysis
- Text Analysis

Used Application of Big Data

Areas	Big Data applications
Targeting customers	Big Data helps understanding customers and target them in personalized fashion.
Science and Research	Big Data helps make machines smarter. For example, Google's Self-driving cars
Security	Big Data is used to keep track of the terrorists and anti-national agencies
Finance	Big Data algorithms are used to analyze market and trading opportunities

CONCLUSION:

Online frauds in the form of phishing and pharming are becoming common. In Phishing, the hacker fraudulently presents oneself online as a trusted authority to trick consumers into giving their personal financial information for identity theft. It is usually perpetuated using mass distributions of e-mails. Pharming is an activity that takes advantage of the vulnerability in Domain Name Service (DNS) server software allowing a hacker to redirect a website's traffic to another web site Big data tools are being used to fight cyber attacks. Cyber crime experts are leveraging big data tools to identify the potential threats and prevent cyber crime incidents. Organizations need not go into complexities of how these cyber-attacks are carried out from an in-depth technical perspective. However, they must concentrate their energies to understand how the attacks break their defense systems. Fighting Cyber crime with actionable insights and data analytics can lead to enhanced information security capabilities. Improved cyber security resilience and reduced business impact is vital as it replaces the idea of reacting to security breaches with requirement process of anticipating the potential risks, evaluating them, and responding in a desired manner.

REFERENCES:

1. Raghupathi W: Data Mining in Health Care. In Healthcare Informatics: Improving Efficiency and Productivity. Edited by Kudyba S. Taylor & Francis; 2010:211–223.
2. Burghard C: Big Data and Analytics Key to Accountable Care Success. IDC Health Insights; 2012.
3. Dembosky A: "Data Prescription for Better Healthcare." Financial Times, December 12, 2012, p. 19; 2012. Available from: http://www.ft.com/intl/cms/ s/2/55cbca5a-4333-11e2-aa8f-00144feabdc0.html#axzz2W9cuwajK.
4. Feldman B, Martin EM, Skotnes T: "Big Data in Healthcare Hype and Hope." October 2012. Dr. Bonnie 360; 2012. http://www.west-info.eu/files/big-data-inhealthcare.pdf.
5. Fernandes L, O'Connor M, Weaver V: Big data, bigger outcomes. J AHIMA 2012:38–42.
6. IHTT: Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry; 2013. http://ihealthtran.com/ wordpress/2013/03/iht%C2%B2-releases-big-data-research-reportdownload-today/.
7. Frost & Sullivan: Drowning in Big Data? Reducing Information Technology Complexities and Costs for Healthcare Organizations. http://www.emc.com/ collateral/analyst-reports/frost-sullivan-reducing-information-technologycomplexities-ar.pdf.
8. Bian J, Topaloglu U, Yu F, Yu F: Towards Large-scale Twitter Mining for Drugrelated Adverse Events. Maui, Hawaii: SHB; 2012.
9. Raghupathi W, Raghupathi V: An Overview of Health Analytics. Working paper; 2013.
10. Ikanow: Data Analytics for Healthcare: Creating Understanding from Big Data. http://info.ikanow.com/Portals/163225/docs/data-analytics-for-healthcare.pdf.
11. IBM: Harvard Medical School; 2011. http://public.dhe.ibm.com/common/ssi/ecm/en/imc14685usen/IMC14685USEN.PDF.
12. Raghupathi W, Kesh S: Interoperable electronic health records design: towards a service-oriented architecture. e-Service Journal 2007, 5:39–57. 28. Borkar VR, Carey MJ, Chen L: Big data platforms: what's next? ACM Crossroads 2012, 19(1):44–49
13. jStart: "How Big Data Analytics Reduced Medicaid Re-admissions." A jStart Case Study; 2012. http://www-01.ibm.com/software/ebusiness/jstart/portfolio/uncMedicaidCaseStudy.pdf. 12. Knowledgent: Big Data and Healthcare Payers; 2013. http://knowledgent. com/mediapage/insights/whitepaper/482.
14. Capgemini: The Deciding Factor: Big Data & Decision Making; 2013. http://www.capgemini.com/thought-leadership/the-deciding-factor-big-datadecision-making.